DEPARTMENT OF APPLIED ECONOMICS

RESEARCH REPORT

PERIODIC REVIEW BASE-STOCK REPLENISHMENT POLICY WITH ENDOGENOUS LEAD TIMES

ROBERT BOUTE • MARC LAMBRECHT • BENNY VAN HOUDT

OR 0448

Periodic Review Base-Stock Replenishment Policy with Endogenous Lead Times

Boute, Robert[1] Lambrecht, Marc[1]
Van Houdt, Benny[2]

[1] *Dept. of Applied Economics, K.U.Leuven*
Naamsestraat 69, 3000 Leuven, Belgium
<first name>.<last name>@econ.kuleuven.ac.be

[2] *Dept. of Mathematics and Computer Science, University of Antwerp*
Middelheimlaan 1, 2020 Antwerpen, Belgium
benny.vanhoudt@ua.ac.be

Abstract

In this paper, we consider a two stage supply chain where the retailer's inventory is controlled by the periodic review, base-stock level (R,S) replenishment policy and the replenishment lead times are endogenously generated by the manufacturer's production system with finite capacity. We extend the work of Benjaafar and Kim (2004) who study the effect of demand variability in a continuously reviewed base-stock policy with single unit demands. In our analysis, we allow for demand in batches of variable size, which is a common setting in supply chains. A procedure is developed using matrix analytic methods to provide an exact calculation of the lead time distribution, which enables the computation of the distribution of lead time demand and consequently the safety stock in an exact way instead of using approximations. Treating the lead time as an endogenous stochastic variable has a substantial impact on safety stock. We numerically show that the exogenous lead time assumption may dramatically degrade customer service.

Keywords: *Production/inventory systems, base-stock replenishment policy, endogenous lead times, safety stock, D-BMAP queueing system, matrix-analytic methods*

1 Introduction

A frequently used stochastic inventory control system is the (R, S) replenishment policy in which every R units of time a replenishment order is placed of sufficient magnitude to raise the inventory position to a base-stock level S. It is referred to as the periodic review, base-stock level control system (Silver et al., 1998). When demand is probabilistic, there is a definite chance of not being able to satisfy some of the demand directly out of stock. Therefore, a buffer or safety stock is required to meet unexpected fluctuations in demand.

Basically, there are two perspectives on how to determine the amount of safety stock. A common approach involves specifying (explicitly of implicitly) a way of costing a shortage and then minimizing total cost. Holding more inventory reduces the probability of a stock out, but increases the inventory holding cost. The cost-minimization approach trades off these costs to find the lowest cost policy. As an alternative, service level requirements are widely used. The service level is a measure of performance for meeting demand from inventory. It can be expressed as the probability of meeting demand from inventory (customer service level) or as the fraction of demand that is met from inventory (fill rate). The service level becomes a constraint in establishing the safety stock of an item; for example, minimize the carrying cost of an item subject to satisfying, routinely from stock, 95 percent of all demands (Silver et al., 1998).

A considerable amount of research is devoted to the determination of the safety stock that is required to ensure a certain service level (Zipkin, 2000). Since the safety stock acts as a buffer between unexpected changes in demand during the replenishment lead time, it is affected by the demand variability, the replenishment lead time, and the lead time variability.

Demand variability is a central theme in the inventory literature and it is generally accepted that higher demand variability degrades service performance (Zipkin, 2000). In particular, Lu et al. (2003) consider an assemble-to-order system where the supply system is modeled using an infinite server queue with compound Poisson demand. They show that increased variability in order sizes degrades order fill rates. Another study done by Jemaï and Karaesmen (2004) analyses a make-to-stock queue and demonstrates that an increase in demand interarrival time variability leads to higher base-stock levels and higher costs.

The effect of lead times is also widely analysed. Song (1994a; 1994b) proves that increased lead time variability causes an increase in the optimal base-stock levels and the optimal costs. More recently, Song and Yao (2002) show that increased lead time variability also degrades fill rate performance. Chopra et al. (2004) state that decreasing lead time is the right lever to cut inventories, rather than reducing lead time variability. Finally, an interesting contribution is done

by Vendemia et al. (1995), who examine the impact of decreasing lead times in conjunction with the characteristics of the demand process. Since inventory variance increases with longer lead times because more random variables are involved, their results indicate that lead time reduction leads to lower costs.

In the inventory literature, the focus has been on inventory systems with exogenous lead times. However, a two stage supply chain can be modelled as a production-inventory system, where the retailer's inventory replenishment lead times are endogenously determined and depend on the current level of congestion in the manufacturer's production system. This is done by Thonemann and Bradley (2002) who consider a decentralized system consisting of a manufacturer and several retailers. They use a queueing model to approximate expected manufacturing lead time, which they then use to approximate lead time demand. By treating leadtimes as i.i.d. random variables, they are able to decouple the analysis of the inventory and production systems.

It is essential to extend pure inventory sytems with exogenous lead times to production-inventory sytems with endogenous lead times. After all, inventory influences production by initiating orders, and production influences inventory by completing and delivering orders to inventory. Therefore the inventory control system should work with a lead time which is a good estimate of the real lead time, depending on the production load, the interarrival rate of orders, the variability of the production system, etc. (Axsäter, 1976). Production-inventory systems with endogenous lead times are also used by Kim and Benjaafar (2002), Benjaafar et al. (2004) and Karaesmen et al. (2002; 2004) to study the effect of respectively inventory pooling, product variety and advance demand information and by Song et al. (1999) to analyse order fulfillment performance measures.

It is the purpose of this paper to study the impact of demand variability on the retailer's safety stock when his inventory is controlled by the periodic review base-stock replenishment policy and where the replenishment lead times are endogenously generated by the manufacturer's production facility with finite capacity. In our setting we also allow for demand in batches. Our main contribution lies in the development of a procedure to estimate the lead time distribution as a function of customer demand, so that the safety stock can be computed in a correct and accurate way. The remainder of the paper is organized as follows. In section 2, we describe our model and the periodic review base-stock policy. In section 3, we explain the procedure that we developed to estimate the lead time distribution based on a queueing analysis. Section 4 analyses the resulting impact on customer service and safety stock. A numerical experiment that illustrates our findings is presented in section 5 and the concluding remarks are given in section 6.

2 Model description

We consider a two stage supply chain with a single retailer and a single manufacturer. Every period, the retailer observes customer demand, denoted by D_t, which represents a finite number of items that customers buy from the retailer. We assume that customer demand D_t is identically and independently distributed (i.i.d.) over time. If there is enough on-hand inventory available, the demand is immediately satisfied. If not, the shortage is backlogged.

In order to replenish the inventory on hand, the retailer places a replenishment order with the manufacturer at the end of every review period. The order quantity O_t is determined by the retailer's replenishment policy. We assume that the manufacturer produces on a make-to-order basis, so he has no finished products inventory. The replenishment orders of size O_t enter the production facility where they are processed on a first-come-first-served basis. When the production facility is busy, they join the queue of unprocessed orders. We assume that the production times for a single product are i.i.d. random variables and to ensure stability, we assume that the utilization of the production facility (average batch production time divided by average batch interarrival time) is strictly smaller than one.

Once the complete batch (equal to the replenishment order) is produced, it is immediately sent to the retailer. The time from the instant the order arrived at the production system to the point that the production of the entire batch is finished, is the response time denoted by T_r. Note that the response time T_r is a continuous variable. In our model however, events occur on a discrete time basis. Therefore we rely on the sequence of events in a period. We assume that the retailer first receives goods from the manufacturer, then he observes and satisfies customer demand and finally, he places a replenishment order with the manufacturer. In this sequence of events, the retailer is always able to satisfy demand after receipt of the products from the manufacturer. Therefore, we round the response time T_r to the smaller integer T_p (i.e., we set $T_p = \lfloor T_r \rfloor$), such that T_p represents the replenishment lead time. A schematic of the system is shown in figure 1.

There are many different types of replenishment policies, of which two are commonly used: on the one hand the periodic review, replenishment interval, base-stock policy and on the other hand the continuous review, reorder point, order quantity model. Given the common practice in retailing to replenish inventories frequently (e.g. daily) and the tendency of manufacturers to produce to demand, we will focus our analysis on the periodic review base-stock replenishment policy.

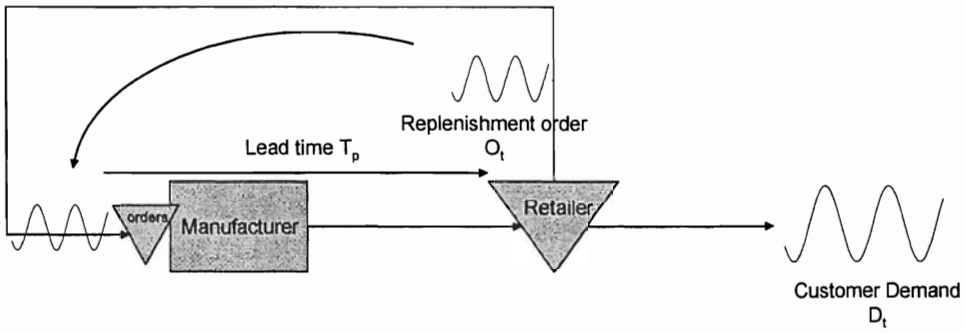

Figure 1: A two stage supply chain modelled as a production-inventory system

In a periodic review base-stock replenishment policy, the retailer tracks his inventory position at the end of every review period R. The inventory position is the sum of the inventory on hand (i.e., items immediately available to meet demand) and the inventory on order (i.e., items ordered but not yet arrived due to the lead time) minus the backlog (i.e., demand that could not be fulfilled and still has to be delivered). A replenishment order is placed to raise the inventory position to a base-stock level S, which determines the order quantity O_t:

$$O_t = S - \textit{inventory position}_t. \tag{1}$$

The base-stock level S is the inventory needed to ensure a given customer service. The risk period is the time between the moment a replenishment order is placed until the subsequent replenishment arrives (i.e., the time from ordering replenishment i to the arrival of replenishment $i+1$) and is equal to the review period plus the lead time $(R + T_p)$. Since the lead time and the demand size are both random i.i.d. variables independent of each other, the demand during this risk period is a sum of a random number of random i.i.d. variables (Ross, 1983). Consequently, the base-stock level equals

$$S = (E[T_p] + R) \times \overline{D} + SS, \tag{2}$$

with $\overline{D}$ the average demand and SS denoting the safety stock.

In our model we do not assume any order or setup costs. Therefore we simple set the review period equal to one base period $(R = 1)$, so that we place a replenishment order every period t. The inventory position at the end of period t is equal to last period's inventory position (which is raised up to the base-stock level S) minus the observed demand in the current period. Hence,

the order quantity placed at the end of period t equals

$$\begin{aligned} O_t &= S - (S - D_t) \\ &= D_t. \end{aligned} \tag{3}$$

We simply order what the demand is in the base period. That is why this policy is also called 'passing on orders' or 'chasing sales'. In the remainder of this paper we will assume the review period R to be one base period. Note however that our framework can similarly be used for review periods $R > 1$.

3 Estimation of the lead time

Most inventory models proposed in the literature take the replenishment lead time T_p as a fixed constant or as an exogenous variable with a given probability distribution. However, the replenishment orders in fact load the production facilities, and the nature of this loading process relative to available capacity and the variability it creates are the primary determinants of lead times in the facility (Karmarkar, 1993). In this section, we develop a procedure to estimate the distribution of the lead time based on a queueing analysis.

The arrival process of orders at the production facility is governed by the retailer's replenishment orders. The periodic review base-stock policy generates batch arrivals with a fixed interarrival time (equal to the review period $R = 1$) and with a variable batch size. In the previous section, it is shown that when the review period is one base period and customer demand is i.i.d., the replenishment orders are exactly the same as the customer demand and hence the batch sizes are i.i.d. random variables with the same distribution as customer demand.

In our procedure, we fit the probability distribution of the order size O_t by a *discrete phase type (PH) distribution.* A discrete PH distribution is the distribution of the number of steps prior to final absorption in an absorbing Markov chain. The key idea behind PH distributions is to exploit the Markovian structure of the distribution to simplify the queueing analysis. Moreover, any general discrete distribution can be approximated in sufficient detail by means of a PH distribution (Horváth and Telek, 2002), since the class of discrete PH distributions is a versatile set that is dense within the set of all discrete distributions on the nonnegative integers (Neuts, 1989; Latouche and Ramaswami, 1999; Bobbio et al., 2003). As we want to analyse the effect of demand variability on customer service, we restrict ourselves to fitting the first two moments only (the mean $\overline{O}$ and standard deviation σ_O). Including more moments (or including a whole empirical distribution) increases the number of phases of the PH distribution and consequently the computational complexity of our algorithm.

We also fit a PH distribution to the mean $\overline{M}$ and standard deviation σ_M of the production time of a single product. Although fitting the mean and standard deviation of the order size O or the production time M by a PH distribution seems analogue, the choice of the time unit of the queueing system of interest offers an additional degree of freedom when fitting the production time M (Bobbio et al., 2004b).

In Step (I) we explain the PH fitting procedure. In Step (II) we develop an efficient algorithm that computes the lead time distribution given the PH distributions obtained in Step (I) by means of Matrix Analytic Methods (MAMs). Matrix analytic techniques, pioneered by Marcel Neuts (1989; 1981), provide a framework that is widely used for the exact analysis of a general and frequently encountered class of queueing models. In these models, the embedded Markov chains are two-dimensional generalizations of elementary GI/M/1 and M/G/1 queues (Kleinrock, 1975), and their intersection, i.e., quasi-birth-death (QBD) processes. We illustrate both steps with an example.

Step (I): Throughout this section we assume that $\overline{O} \geq 2$ is an integer; this condition is not necessary, but allows some simplification in the fitting procedure. A PH distribution X is characterized by the triple (n, T, α), where $n > 0$ is an integer, referred to as the number of phases of the distribution or the number of transient states in an absorbing Markov chain, T is an $n \times n$ substochastic matrix, delineating the transition probabilities between the transient states and α is a stochastic $1 \times n$ vector, which defines the probabilities α_i that the process is started in the transient state i. The transition probabilities between the transient states and the absorbing state are given by t, which is an $n \times 1$ substochastic vector equal to $e - Te$, where e is a $n \times 1$ column vector with all its entries equal to one. Hence the probability that k steps are taken prior to absorption is given by

$$\Pr[X = k] = \alpha T^{k-1} t, \tag{4}$$

where $k > 0$. Its mean $\overline{X}$ and standard deviation σ_X obey the following equations:

$$\overline{X} = \alpha (I - T)^{-1} e, \tag{5}$$

$$\sigma_X = \sqrt{\alpha (I - T)^{-1} \left(2(I - T)^{-1} + (1 - \overline{X}) I\right) e}, \tag{6}$$

with I an $n \times n$ identity matrix. In order to match the mean order size $\overline{O}$ and its standard deviation σ_O, we need to find a PH distribution characterized by a triple (n, T, α) such that $\overline{O} = \overline{X}$ and $\sigma_O = \sigma_X$. Moreover, since the algorithm

developed in Step (II) speeds up with a smaller n, we want a representation (n, T, α) that fits the two first moments with n as small as possible (including higher moments will lead to a higher n).

Denote cv_O^2 as the squared coefficient of variation, that is, $cv_O^2 = \sigma_O^2/\overline{O}^2$. By applying a theorem by Telek (2000, Theorem 1) and the fact that $\overline{O}$ is an integer, we find that the minimum number of phases needed to match $\overline{O}$ and σ_O equals

$$n = \max\left(2, \left\lceil \frac{\overline{O}}{\overline{O}cv_O^2 + 1} \right\rceil\right). \tag{7}$$

Next, we choose the $1 \times n$ vector α and the $n \times n$ matrix T as follows. These choices are motivated by a variety of results when matching continuous time PH distributions (Telek and Heindl, 2002; Bobbio et al., 2004a):

$$\alpha = (\beta, 1-\beta, 0, 0, \ldots, 0), \tag{8}$$

$$T = \begin{bmatrix} 1-p_1 & p_1 & 0 & 0 & \ldots & 0 \\ 0 & 1-p_2 & p_2 & 0 & \ldots & 0 \\ 0 & 0 & 1-p_2 & p_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \ddots & p_2 \\ 0 & 0 & 0 & 0 & \ldots & 1-p_2 \end{bmatrix}. \tag{9}$$

This leaves us with 3 parameters: β, p_1 and p_2, and two equations: $\overline{O} = \overline{X}$ and $\sigma_O = \sigma_X$. Therefore, we add an additional constraint which demands that the stationary vector of the matrix $T + t\alpha$ is the uniform vector $(1/n, \ldots, 1/n)$. When combined with the requirement $\overline{O} = \overline{X}$, this poses the following conditions on β, p_1 and p_2:

$$\begin{aligned} p_1 &= \beta n/\overline{O}, \\ p_2 &= n/\overline{O}. \end{aligned} \tag{10}$$

Remark, $0 \leq p_2 \leq 1$ and $0 \leq p_1 \leq \beta$ as $\overline{O} \geq n$. Thus, it remains to determine β, with $0 \leq \beta \leq 1$, based on the remaining condition $\sigma_O = \sigma_X$.

Let $G(z) = \sum_k P[X = k] z^k$ be the generating function of the PH distribution X characterized by Eqns. (8) and (9). Then,

$$G(z) = \left\{\beta\left(\frac{p_1 z}{1-(1-p_1)z}\right) + (1-\beta)\right\}\left(\frac{p_2 z}{1-(1-p_2)z}\right)^{n-1}, \tag{11}$$

and the condition $\sigma_O = \sigma_X$ can be rephrased as

$$\sigma_O^2 = \left.\frac{d^2G(z)}{dz^2}\right|_{z=1} + \overline{O}(1-\overline{O}). \tag{12}$$

Some careful calculations show that this equation is solved by setting β equal to:

$$0 \leq \beta = \frac{2\overline{O}}{2\overline{O} + n(n - \overline{O} + n\ cv_O^2\overline{O})} \leq 1, \tag{13}$$

where the first and last inequality is due to Eqn. (7). In conclusion, by making use of Eqns. (7-10) and (13), we can fit the mean $\overline{O}$ and standard deviation σ_O by a PH distribution.

The same procedure can be used to match the mean $\overline{M}$ and standard deviation σ_M of the production time of a single product (if we replace the necessary O's into M's). However, in this case we can do significantly better. Since the lead time is expressed as an integer number of periods and the interarrival time is equal to one base period, we have the freedom to choose the time unit U of the queueing system in an appropriate manner (Bobbio et al., 2004b). Let U equal half of the mean production time of an item, i.e., $U = \overline{M}/2$, and denote $\overline{M}_U$ and σ_{M_U} as the mean and standard deviation of the production time expressed as multiples of U. By definition, we find $\overline{M}_U = 2$ and $\sigma_{M_U} = 2\sigma_M/\overline{M}$, implying that $cv_M^2(U) = cv_M^2$. Consequently, we only need $n = 2$ phases, because

$$n = \max\left(2, \left\lceil \frac{2}{1 + 2cv_M^2} \right\rceil\right) = 2. \tag{14}$$

Meaning, we can always match the production process of a single product using a 2 state PH distribution. The remainder of the matching algorithm is identical to the procedure used to fit the order size distribution.

Example

- Suppose that every week an order arrives at the manufacturer's production facility with an average size of 50 products and a standard deviation of 25 products. Our objective is to fit a PH distribution to

$$\begin{cases} \overline{O} = 50 \\ \sigma_O = 25, \end{cases}$$

with n as small as possible (as this will speed up our algorithm). Applying

Eqns. (7), (13) and (10) successively gives

$$n_O = \max\left(2, \left\lceil \frac{50}{50 \cdot (0.25) + 1} \right\rceil\right) = 4,$$

$$\beta = \frac{2 \cdot 50}{2 \cdot 50 + 4\,(4 - 50 + 4 \cdot (0.25) \cdot 50)} = 0.8261,$$

and

$$\begin{aligned} p_1 &= \frac{(0.8261) \cdot 4}{50} = 0.069, \\ p_2 &= \frac{4}{50} = 0.08. \end{aligned}$$

According to Eqns. (8) and (9), this results in

$$\alpha_O = (0.8261, 0.1379, 0, 0),$$

and

$$T_O = \begin{bmatrix} 0.931 & 0.069 & 0 & 0 \\ 0 & 0.92 & 0.08 & 0 \\ 0 & 0 & 0.92 & 0.08 \\ 0 & 0 & 0 & 0.92 \end{bmatrix}.$$

The resulting discrete PH distribution characterized by the triple (n_O, T_O, α_O) describes the distribution of the number of steps prior to final absorption in an absorbing Markov chain with 4 transient states and 1 absorbing state. The probability that the process is started in a transient state $i = 1, 2, 3, 4$, is given by $\alpha_O = (0.8261, 0.1379, 0, 0)$, and the 5×5 state transition matrix of the Markov chain is given by

$$\begin{bmatrix} T_O & t_O \\ 0 & 1 \end{bmatrix},$$

with T_O delineating the transition probabilities between the transient states and t_O the transition probabilities between the transient states and the absorbing state ($t_O = e - T_O e$):

$$t_O = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.08 \end{bmatrix}.$$

Hence the probability that the order size is equal to e.g. 20 corresponds to the probability that the number of steps from the moment the process is started until it eventually absorbs in the absorbing state is equal to 20 and can be found from

$$\begin{aligned} \Pr[O = 20] &= \alpha_O (T_O)^{19} t_O \\ &= 0.0111. \end{aligned}$$

The complete discrete PH distribution of the order pattern is plotted in figure (2).

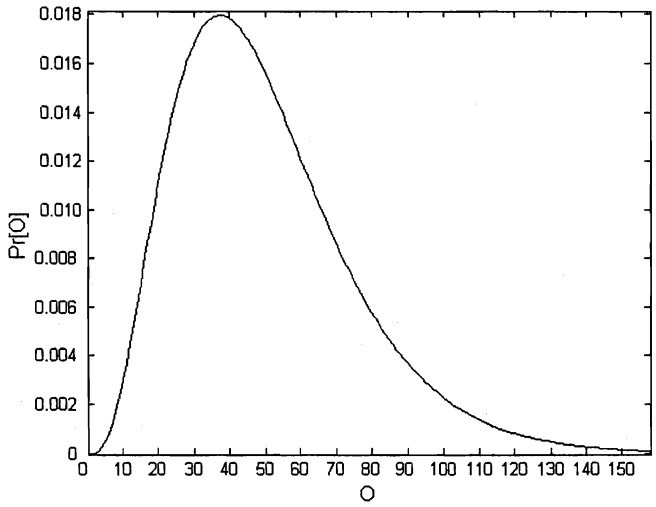

Figure 2: Discrete PH type distributed order process fitted to $\overline{O} = 50$ and $\sigma_O = 25$.

- Assume that the production facility is available ten hours per day, five days a week. It takes on average 54.054 minutes to produce a single product with a coefficient of variation equal to one. Since orders arrive with an average batch size of 50 products per week, the production facility is on average 50×54.054 minutes busy per week, which results in an average load of $2702.7/3000 = 90.09\%$. We first choose the time unit of our queueing system equal to $U = \overline{M}/2 = 54.054/6000$, so that we actually fit a PH distribution to

$$\begin{cases} \overline{M}_U = 2 \\ \sigma_{M_U} = 2, \end{cases}$$

which denote the mean and standard deviation of the single product service times expressed in time units U. Applying the same procedure results in the triple (n_M, T_M, α_M) given by

$$n_M = 2,$$

$$\alpha_M = (0.3333, 0.6667),$$

$$T_M = \begin{bmatrix} 0.6667 & 0.3333 \\ 0 & 0 \end{bmatrix}.$$

Step (II): Note that in Step (I) we chose U as the time unit of our queueing system. As a consequence, interarrival times, production times and response times are expressed in this time unit U. For instance, a replenishment order placed every period ($R = 1$) is translated into a new batch arrival occuring every d time slots, with $d = 1/U$.

Let (n_O, T_O, α_O) and (n_M, T_M, α_M) be the matrix representations of both PH distributions obtained in Step (I). We can reduce the complexity of our queueing analysis by exploiting the phase-type nature of the fitted order size distribution. To do so, we first apply a discrete time variant of one of the closure properties of PH distributions (Latouche and Ramaswami, 1999, Theorem 2.6.3), to find that the production time of an entire batch (an integral replenishment order) is PH distributed with representation (n_S, T_S, α_S):

$$\begin{aligned} n_S &= n_O n_M = 2n_O, \\ T_S &= (I_{n_O} \otimes T_M) + (T_O \otimes t_M \alpha_M), \\ \alpha_S &= \alpha_O \otimes \alpha_M, \end{aligned} \tag{15}$$

where $\otimes$ denotes the Kronecker product between matrices, $t_M = e - T_M e$ and I_x is an identity matrix of dimension x.

This permits us to treat the entire batch order as a single customer and hence, the problem of estimating the lead time is reduced to computing the response time distribution of a customer in a $D/PH/1$ queue. Such a queue has a deterministic arrival process that generates a single customer (equal to a replenishment order with batch size O) at fixed time epochs, i.e. every d time units. The single service center serves this customer in a PH distributed time characterized by the triple (n_S, T_S, α_S). The size of the waiting line is infinite (meaning, we assume that orders are never lost nor rejected by the manufacturer).

To compute the response time T_r, we construct a Markov chain (MC) (A_n, S_n), where A_n represents the age of the order in service at the n-th observation point t_n and S_n reflects the phase of the service process at epoch t_n. The age of the order in service at time t_n is defined as the duration (expressed in time slots) of the time interval $[a_n, t_n)$, where a_n denotes the arrival time of the replenishment order. Instead of observing the Markov chain (A_n, S_n) at all time epochs, we observe the system only when the server is busy (simplifying the boundary behavior of the Markov chain). All events such as arrivals, transfers from the waiting line to the server and service completions are assumed to occur at instants immediately after the discrete time epochs. This implies, amongst others, that the age of an order in service at some time epoch t_n is at least 1.

The MC (A_n, S_n) has an infinite number of states labeled $1, 2, \ldots$. The set of states $\{(i-1)n_S + 1, \ldots, in_S\}$ is referred to as level i of the MC, for $i \geq 1$. The states of level $i > 0$ are labeled as s, where $1 \leq s \leq n_S$. Let state s of level i of the MC correspond to the situation in which there is an order in service (being produced), that arrived i time units ago, while the service process is currently in phase s. The transition matrix P of this MC can be written as

$$P = \begin{bmatrix} A_d & A_0 & 0 & \ldots & 0 & 0 & \ldots \\ A_d & 0 & A_0 & \ldots & 0 & 0 & \ldots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots & \ddots \\ A_d & 0 & 0 & \ldots & A_0 & 0 & \ldots \\ 0 & A_d & 0 & \ldots & 0 & A_0 & \ddots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \ddots \end{bmatrix}, \tag{16}$$

where $A_0 = T_S$ and $A_d = t_S \alpha_S$ (and $t_S = e - T_S e$). Let us explain this by distinguishing between two cases:

(i) Assume that a batch order of age $q \geq d$ is in service at time t_n and the service process is in phase s (recall, d is the interarrival time). Then, either (a) his service is completed at time t_n, with probability $(t_S)_s$, or (b) his service continues and the new phase of the service process equals s', with probability $(T_S)_{s,s'}$. In scenario (a), the next order, who arrived at time $a_{n+1} = a_n + d$, has an age $q' = q - d + 1 \geq 1$ at time $t_{n+1} = t_n + 1$, as $q' = t_{n+1} - a_{n+1}$ and $q = t_n - a_n$. The new phase s' of the service process is determined by α_S. In scenario (b), the same order remains in service, hence, $t_{n+1} = t_n + 1$ and $q' = q + 1$.

(ii) Assume the age q of the order in service is less than d at time t_n. The scenario where this order remains in service is identical to (i). However, if a service completion occurs, the next order arrives at time $a_{n+1} = a_n + d =$

$t_n - q + d > t_n$. Since all arrivals occur immediately after the discrete time epochs, the service facility is empty at the time instants $t_n + 1, \ldots, t_n + d - q$. Therefore, $t_{n+1} = t_n + d - q + 1$ and the new age $q' = t_{n+1} - a_{n+1} = (t_n + d - q + 1) - (t_n - q + d) = 1$, meaning the Markov chain makes a transition to level 1. The new phase s' is, once more, determined by the vector α_S.

The MC characterized by Eqn. (16) is of the GI/M/1 type (Neuts, 1981). Using Neuts' stability condition, one easily finds that this MC is ergodic if and only if $\rho < 1$, i.e., $\bar{O}\overline{M}_U < d$. For an ergodic MC of the GI/M/1 type, one computes the steady state vector π of P, that is, $\pi P = \pi$ and $\pi e = 1$, as follows:

$$\pi_1 = \alpha_S / (\alpha_S (I - T_S)^{-1} e), \tag{17}$$

$$\pi_i = \pi_1 Y^{i-1}, \tag{18}$$

where $\pi = (\pi_1, \pi_2, \ldots)$ and π_i is a $1 \times n_S$ vector, for all $i > 0$. The expression for π_1 is obtained using the normalization condition $\sum_i \pi_i e = 1$ and the stochastic interpretation of π_1. The $n_S \times n_S$ rate matrix Y is the smallest nonnegative solution to the matrix equation $Y = A_0 + Y^d A_d$ and can be numerically solved with a variety of algorithms, e.g. Neuts (1981), Ramaswami (1988) and Alfa et al. (2002).

Having obtained the steady state vector $\pi = (\pi_1, \pi_2, \ldots)$, we can obtain the response time T_r using the following observation: The probability that a batch order has a response time of i time units can be calculated as the expected number of orders with an age of i time units that complete their service at an arbitrary time instant, divided by the expected number of orders that complete their service during an arbitrary time instant (that is, $1/d$ for a queue with $\rho < 1$). As such, we find the response time distribution as

$$\Pr[T_r = i] = d\rho\pi_i t_S. \tag{19}$$

Notice, the s-th element of $\rho\pi_i$ equals the probability that an order of age i is in service at an arbitrary time instant with the service process in phase s.

In step (I) we chose the time unit U of our queueing system as half of the production time of a single product (i.e., $\overline{M}/2$). Thus, if we want to express the replenishment lead time in terms of the number of periods needed to deliver the order to the retailer, we still need to make the following conversion:

$$\Pr[T_p = i] = \sum_j \Pr[T_r = j] 1_{\{\lfloor j/d \rfloor = i\}}, \tag{20}$$

where 1_A is 1 if the event A is true and 0 otherwise. Note that this conversion at the same time rounds the (possibly fractional) response time T_r to the discrete replenishment lead time T_p.

Example

We continue with the example described above. Treating the entire batch order as a single customer results in a $D/PH/1$ queue, where the service time of the entire batch is PH distributed since both batch size and single product service times are PH distributed. As the time unit is chosen to be $U = 54.054/6000$ or 27.027 minutes, this queue has a deterministic arrival process with an interarrival time of $d = 6000/54.054 = 111$. The batch production time is characterized by the triple (n_S, T_S, α_S) given by Eqn. (15):

$$n_S = 8,$$

$$\alpha_S = (0.2874, 0.5747, 0.0460, 0.0920, 0, 0, 0, 0)\,,$$

$$T_S = \begin{bmatrix} 0.6667 & 0.3333 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.3103 & 0.6207 & 0.0230 & 0.0460 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.6667 & 0.3333 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.3067 & 0.6133 & 0.0267 & 0.0533 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.6667 & 0.3333 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3067 & 0.6133 & 0.0267 & 0.0533 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.6667 & 0.3333 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.3067 & 0.6133 \end{bmatrix}.$$

To compute the response time of this $D/PH/1$ queue, we construct a Markov Chain (A_n, S_n) with A_n representing the age of the batch in service and S_n the phase of the service process. To see how this works, we take for instance that at time t_n a batch order is already more than a week in the system, e.g. 54.054 hours, and its service process is currently in phase 8. Then, either the production is finished with probability 0.08,

which is given by $t_S = e - T_S e$:

$$t_S = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.08 \end{bmatrix},$$

or its service continues (with probability 0.92).

- In case the batch order is not yet completely produced, then the order's age increases with one time unit U at the next time step t_{n+1} (resulting in an order's age of 54.054 hours + 27.027 minutes or 54.504 hours) and the new phase of the service process is phase 7 with probability 0.3067 or phase 8 with probability 0.6133 (dictated by T_S).

- In case the production is finished, the response time equals 54.054 hours. The next batch order that arrived a week (50 hours) later than the current order is 4.054 hours in the system at the moment that current production finishes, and hence it will be 4.054 hours + 27.027 minutes or 4.504 hours in the system at the next time step in the Markov Chain (t_{n+1}). The new phase of the service process can be $1, 2, 3$ or 4 with respective probabilities of $0.2874, 0.5747, 0.0460$ and 0.0920 (given by α_S). A response time T_r equal to 54.054 hours corresponds to 1.08 weeks. These products are immediately sent to the retailer. Since these products can be used to fulfill the customer demand that is observed within the second week after the order is placed, we round the replenishment lead time T_p to 1 week.

  These transitions constitute the transition matrix P, described by (16). Once we find the response times based on the steady state vector of this transition matrix, we finally convert them to discrete replenishment lead times using Eqn. (20). The resulting discrete lead time distribution is graphically illustrated in figure 3. The average replenishment lead time and its standard deviation are given by

$$\begin{aligned} E[T_p] &= 1.31489 \text{ weeks}, \\ \sigma[T_p] &= 1.3474 \text{ weeks}. \end{aligned}$$

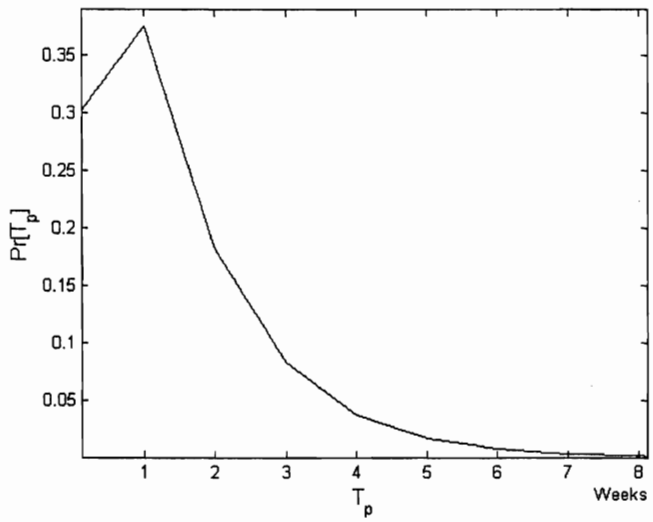

Figure 3: Discrete lead time distribution for $\overline{O} = 50$, $\sigma_O = 25$ and $\overline{M} = \sigma_M = 54.054/3000$ weeks.

4 Impact on safety stock

Managers have been under increasing pressure to decrease inventories as supply chains attempt to become leaner. The goal, however, is to reduce inventories without hurting the level of service provided to customers. For a given service level, a decision maker has three levers that affect safety stock: demand variability, replenishment lead time, and lead time variability (Chopra et al., 2004). In the previous section, we developed a procedure how lead time distribution can explicitly be calculated given the demand pattern. In this section, we focus on the resulting impact on safety stock.

Similar to Graves (1999) and Disney et al. (2004), we characterize the inventory random variable and use it to find the safety stock requirements for the system. After all, the variance of inventory will have an immediate effect on customer service: the higher the inventory variance, the more stock will be needed to maintain customer service at the target level.

The inventory balance can be found as follows. Every period t, the retailer receives the order that he placed $T_p + 1$ periods ago and he satisfies the observed customer demand (D_t) from inventory. Hence

$$NS_t = NS_{t-1} + O_{t-(T_p+1)} - D_t \quad \text{for} \quad t = 1, 2, ..., \tag{21}$$

where NS_t denotes the net stock or inventory on hand at the end of period t.

If $D_t = O_t = 0$ for $t < 0$, then the initial inventory level NS_0 should cover the expected demand until the first replenishment in period $T_p + 1$ plus a buffer or safety stock SS. Hence the initial inventory level equals the base-stock level S. By repeated backward substitution, we can rewrite Eqn. (21) for $t \geq T_p + 1$ as

$$\begin{aligned} NS_t &= SS + (E[T_p] + 1)\overline{D} + \sum_{k=T_p+1}^{t} O_{t-k} - \sum_{k=0}^{t} D_{t-k} \\ &= S + \sum_{k=T_p+1}^{t} O_{t-k} - \sum_{k=0}^{t} D_{t-k}. \end{aligned} \tag{22}$$

In a periodic review base-stock replenishment policy, we place an order equal to the observed customer demand. Hence, substituting Eqn. (3) into the above expression gives

$$NS_t = S - \sum_{k=0}^{T_p} D_{t-k}, \tag{23}$$

with T_p a non-negative integer random variable. The net stock is equal to the initial inventory or base-stock level S minus the demand during the lead time and the review period. From this, we can easily find measures for the average and variance of the net stock. Since T_p is independent of $D_t, D_{t-1}, ...$, the mean and variance of this sum of a random number of random i.i.d. variables is equal to (Ross, 1983):

$$\begin{aligned} E(NS) &= S - (E[T_p] + 1) \times \overline{D} \\ &= SS, \end{aligned} \tag{24}$$

$$\begin{aligned} Var(NS) &= (E[T_p] + 1) \times Var(D) + \overline{D}^2 \times Var(T_p) \\ &= \left(\widehat{\sigma}^{T_p+1}\right)^2. \end{aligned} \tag{25}$$

Our purpose is to find the minimum amount of safety stock that is needed to assure a specified service level. A popular metric to measure customer service is the fill rate, which measures the proportion of the demand that can immediately be delivered from inventory on hand:

$$\text{Fill rate} = 1 - \frac{\text{expected number of backorders}}{\text{expected demand}}.$$

In other words, it is based on the fraction of shortages that occur on average per period when a stockout occurs. From Eqn. (23) we find that the probability of a stockout is the probability that lead time demand exceeds the base-stock level:

$$\Pr\left[NS_t < 0\right] = \Pr\left[\sum_{k=0}^{T_p} D_{t-k} > S\right]. \tag{26}$$

Consequently, the average number of shortages can be written as

$$E\left[NS_t^-\right] = E\left[\left[\sum_{k=0}^{T_p} D_{t-k} - S\right]^+\right], \tag{27}$$

and the fill rate can be calculated as

$$\text{Fill rate} = 1 - \frac{1}{\overline{D}} \cdot E\left[\left[\sum_{k=0}^{T_p} D_{t-k} - S\right]^+\right]. \tag{28}$$

In order to compute the fill rate, we need to determine the distribution of demand during the replenishment lead time and the review period (also called 'lead time demand'). Very often a normal distribution is used to approximate the distribution of lead time demand. However, this may yield significant errors. Bagchi et al. (1986) and Chopra et al. (2004) recommend to use the exact compound distribution of demand during the lead time instead of using approximations. Since we know both the demand and the lead time distribution, we can easily find the distribution of the demand during the replenishment lead time and the review period. Let $D(z)$ and $T_p(z)$ be the probability generating functions of respectively demand and lead time distribution,

$$D(z) = \sum_{n=0}^{\infty} P\left[D = n\right]\ z^n,$$
$$T_p(z) = \sum_{n=0}^{\infty} P\left[T_p = n\right]\ z^n,$$

then the generating function $D_{T_p+1}(z)$ of the demand during lead time and

review period can be written as

$$\begin{aligned} D_{T_p+1}(z) &= \sum_{n=0}^{\infty} P[D^{(T_p+1)*} = n]\ z^n \\ &= \sum_{n=0}^{\infty}\sum_{i=1}^{\infty} P[D^{i*} = n]\ z^n P[T_p + 1 = i] \\ &= D(z) \sum_{i-1=0}^{\infty} P\left[T_p = i-1\right] D(z)^{i-1} \\ &= D(z) \cdot T_p\left(D\left(z\right)\right), \end{aligned}$$

where X^{n*} denotes the n-fold convolution of the random variable X with itself.

For our numerical example described in section 3, we plot the discrete probability distribution of the demand during lead time and review period in figure 4. From this distribution, we can derive its mean and standard deviation:

$$\begin{aligned} E\left[\sum_{k=0}^{T_p} D_{t-k}\right] &= 115.7447, \\ \sigma\left[\sum_{k=0}^{T_p} D_{t-k}\right] &= 77.3648, \end{aligned}$$

which could also directly be found from

$$\begin{aligned} E\left[\sum_{k=0}^{T_p} D_{t-k}\right] &= \left(E\left[T_p\right] + 1\right) \times \overline{D} \\ &= (1.3149 + 1) \times 50, \\ \sigma\left[\sum_{k=0}^{T_p} D_{t-k}\right] &= \sqrt{\left(E\left[T_p\right] + 1\right) \times Var(D) + \overline{D}^2 \times Var(T_p)} \\ &= \sqrt{(1.3149 + 1) \times 25^2 + 50^2 \times (1.3474)^2}. \end{aligned}$$

In practice, decision makers often have to find the minimum safety stock that is required to achieve a given fill rate. Since we know the exact distribution of lead time demand, we can find from Eqn. (28) the minimal base-stock level S that is required such that an imposed fill rate is met. This on his turn results in the optimal safety stock equal to

$$SS = S - \left(E\left[T_p\right] + 1\right)\overline{D}. \tag{29}$$

Suppose that the safety stock should be high enough to satisfy 95% of customer demand immediately from stock. The smallest base-stock level S that

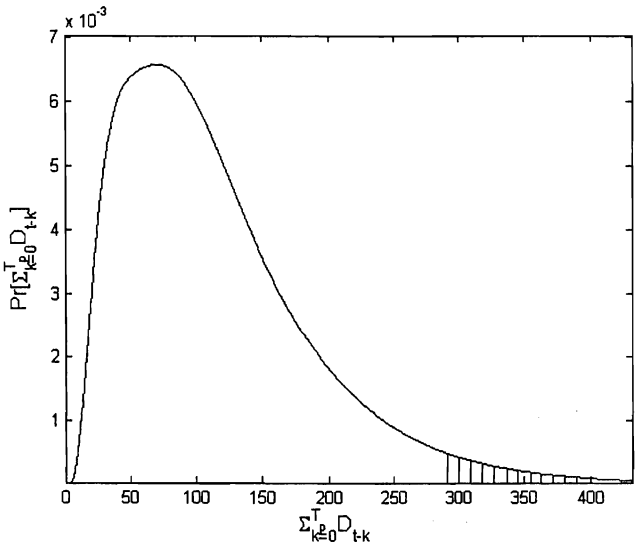

Figure 4: Discrete probability distribution of demand during lead time and review period for $\overline{D} = 50$, $\sigma_D = 25$ and $\overline{M} = \sigma_M = 54.054/3000$ weeks.

achieves this fill rate is equal to

$$S = 289.$$

Hence, the required safety stock should be at least

$$\begin{aligned} SS &= 289 - 115.7447 \\ &= 173.2553, \end{aligned}$$

meaning that a target buffer of 174 products should be kept in inventory.

5 Numerical results

In the previous sections, we described how the lead time distribution can be calculated given the mean and the variance of customer demand and we showed how the safety stock can be exactly computed given the distribution of demand and lead time. In this section, we will present some observations and illustrate with numerical examples.

Observation 1 *A more variable demand pattern generates a longer and a more variable lead time. As congestion increases, this effect is even stronger. As a result, lead time demand becomes considerably larger and more variable*

and consequently, a substantial higher safety stock is required to maintain a given customer service.

This result is intuitively clear. In a periodic review base-stock policy with i.i.d. demand, orders are equal to observed customer demand. As a consequence, a more variable demand pattern implies more variability in the arrival pattern at the production facility, so that queueing performance degrades and lead times increase (Hopp and Spearman, 2001). This on his turn has a reinforced effect on the demand during the lead time, which becomes larger and more variable. Maintaining customer service at a given target level implies that a significant larger amount of safety stock is needed.

We extend the example used in section 3. The retailer observes a variable customer demand with an average of 50 products per week. At the end of the week, the retailer places a replenishment order equal to the observed customer demand. We analyse the effect of the demand variability by changing customer demand's standard deviation over a range from 1 to 50 products per week. Figures 5 and 6 numerically show how the average lead time and its standard deviation increase as customer demand's standard deviation increases when the load is 90%. For example, if variability in customer demand decreases to a standard deviation of only 20 products per week, then the replenishment lead times decrease to an average of 0.97 periods and a standard deviation of 0.96 periods. If, on the contrary, customer demand's variability increases to a standard deviation of 40 products per week, then lead times increase as well to an average of 2.86 periods with a standard deviation of 3.10 periods. In figures 7 and 8 we repeat the same experiment for an average production time of 58.82 minutes (coefficient of variation equal to 1), which corresponds to a congestion of 98.04%. We numerically find that the effect on lead times is even stronger in this case.

In figures 9 and 10, we observe the effect of the demand variability on the mean and standard deviation of the resulting lead time demand for our initial load of 90%. The coefficient of variation of lead time demand is plotted in figure 11. Finally, figure 12 plots the safety stock that is required to achieve a fill rate of 95 percent. It is clear that more variability in customer demand has a substantial impact on the required safety stock. For instance, when customer demand's variability increases to a standard deviation of 40 products per week, then safety stock should amount 555 products in order to achieve a fill rate of 95%. An increase of customer demand's standard deviation to 50 products per week would even give rise to a safety stock of 977 products.

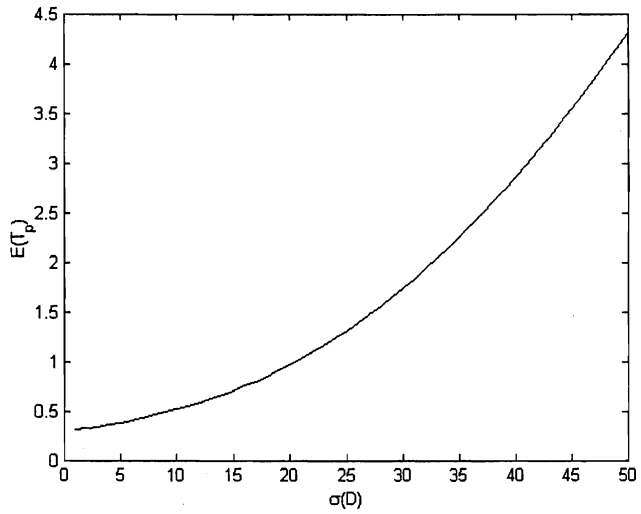

Figure 5: Effect of demand variability on average lead time for $\rho = 0.90$

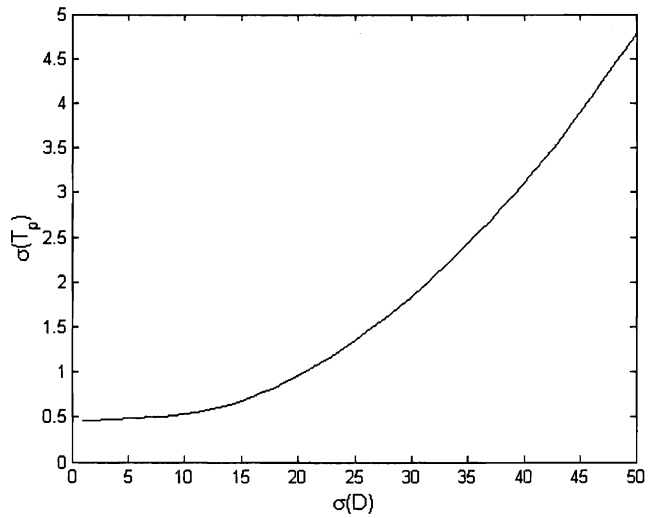

Figure 6: Effect of demand variability on lead time variability for $\rho = 0.90$

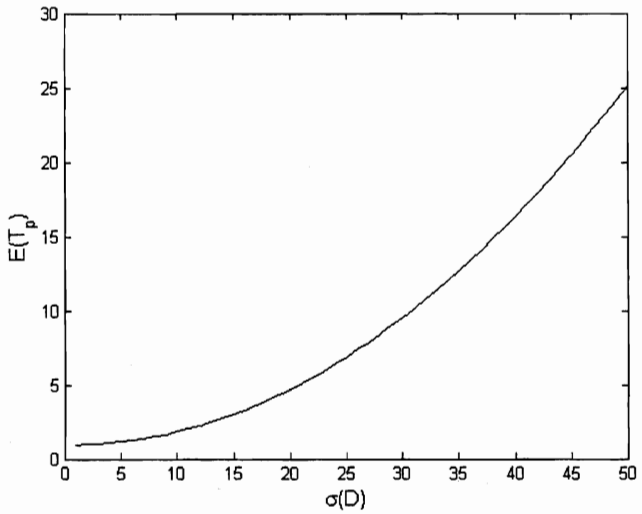

Figure 7: Effect of demand variability on average lead time for $\rho = 0.98$

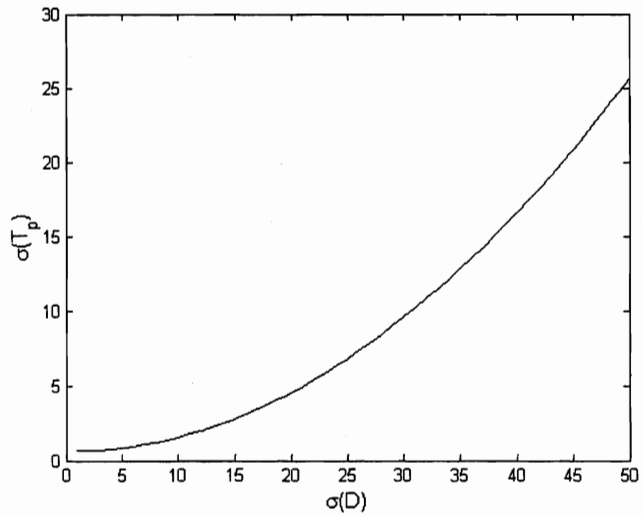

Figure 8: Effect of demand variability on lead time variability for $\rho = 0.98$

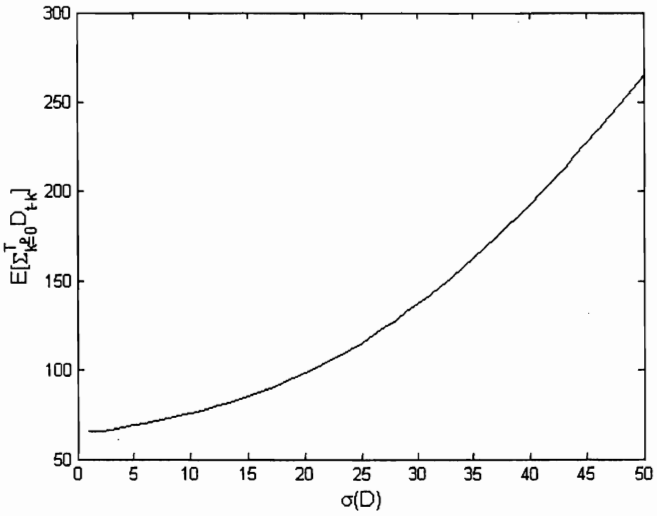

Figure 9: Effect of demand variability on average lead time demand

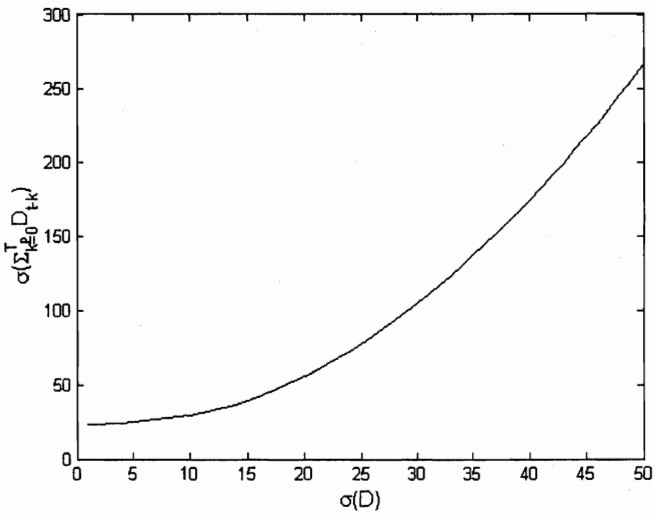

Figure 10: Effect of demand variability on the variability of lead time demand

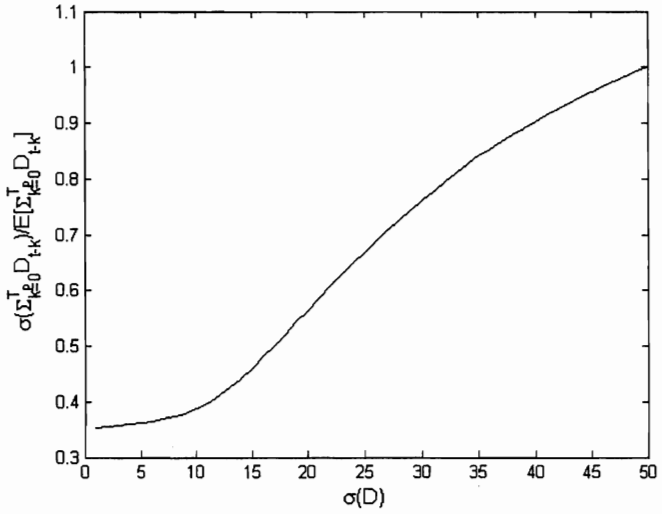

Figure 11: Effect of demand variability on coefficient of variation of lead time demand

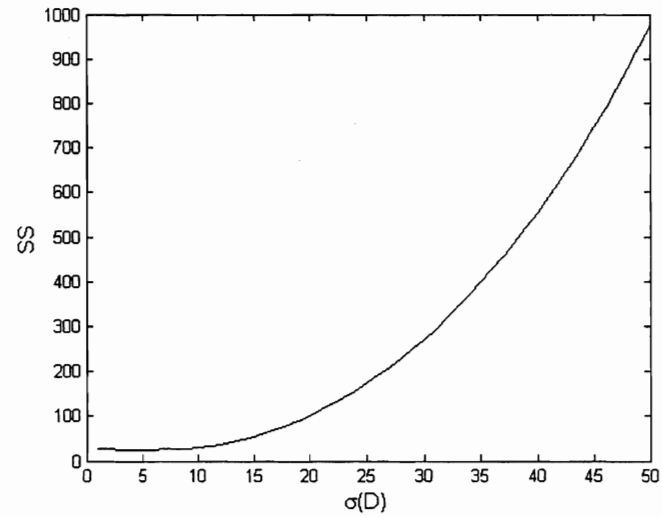

Figure 12: Effect on demand variability on safety stock

Observation 2 *Ignoring endogenous lead times incorrectly calculates safety stocks and may seriously underestimate customer service.*

Assume for instance that the retailer faces a variable customer demand of 50 products per week on average and a standard devation of 25 products. The retailer knows that his lead time distribution is given by figure 3 on which he bases his safety stock decision. The base-stock level S equals 289 and the safety stock 174 products. When customer demand becomes more variable, i.e. standard deviation increases to 40 products per week, and the retailer does not include the effect of this demand variability increase on the lead time distribution in his safety stock calculations (i.e., he still uses the lead time distribution given by figure 3), then he would set a base-stock level equal to $S\prime = 331$ and a corresponding safety stock equal to $SS\prime = 331 - 50 \times (1 + 1.31489) = 215.26$ to achieve a supposed fill rate of 95% (see figure 13).

However, when the retailer holds on this base-stock level and safety stock, he obtains a fill rate of less than 43% (see figure 14) due to the increase of the lead times and lead time variability caused by the increase in demand variability. This demonstrates the importance of including endogenous lead times in inventory control models. The exogenous lead time assumption dramatically degrades customer service.

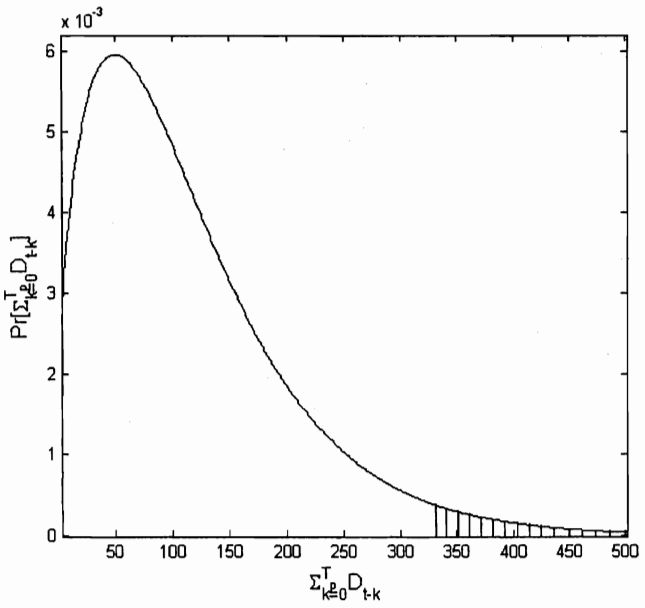

Figure 13: Discrete probability distribution of demand during lead time and review period for $\overline{D} = 50$, $\sigma_D = 40$ with exogenous lead times given by the lead time distribution in figure 3.

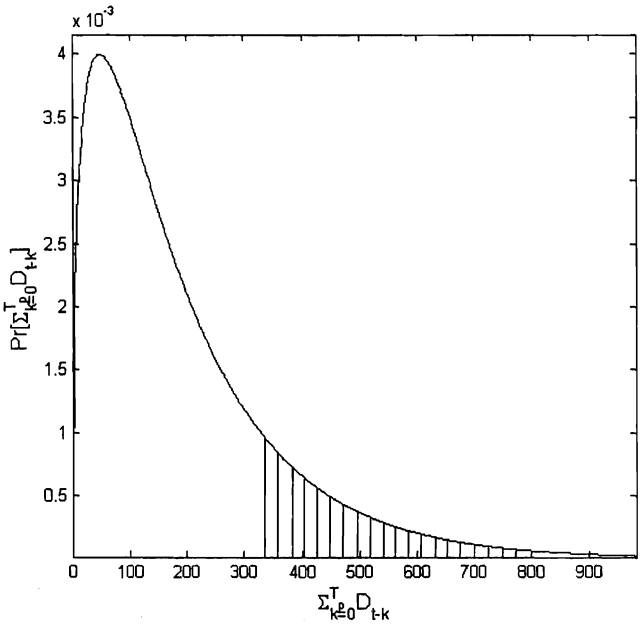

Figure 14: Discrete probability distribution of demand during lead time and review period for $\overline{D} = 50$, $\sigma_D = 40$ with endogenous lead time determination.

6 Conclusions

Under a periodic review base-stock replenishment policy and an i.i.d. customer demand, a more variable demand or a more variable lead time forces the safety stock to increase in order to maintain the customer service at a target level. In this paper, we incorporate the lead time as an endogenous variable and we explicitly analyse the impact of the demand pattern on the lead time distribution. The lead time distribution is obtained using phase type distributions and matrix analytic methods. We numerically observe that in this production-inventory system a more variable demand pattern results in a longer and more variable lead time and due to this reinforced effect, a substantial higher safety stock is required to maintain a given service level. Ignoring endogenous lead times may degrade customer service considerably.

Acknowledgements

This research contribution is supported by contract grant G.0051.03 from the Research Programme of the Fund for Scientific Research – Flanders (Belgium) (F.W.O.-Vlaanderen). Benny Van Houdt is a postdoctoral Fellow of F.W.O.-

Vlaanderen. The authors would like to thank M. Telek for his helpful discussion on the PH fitting procedure.

References

Alfa, A., Sengupta, B., Takine, T. and Xue, J. (2002). A new algorithm for computing the rate matrix of GI/M/1 type Markov chains. *Proc. of the 4th Int. Conf. on Matrix Analytic Methods.* Adelaide, Australia. pp 1–16.

Axsäter, S. (1976). Coordinating control of production-inventory systems. *International Journal of Production Research*, 14(6), pp 669–688.

Bagchi, U., Hayya, J. C. and Chao-Hsien, C. (1986). The effect of lead-time variability: The case of independent demand. *Journal of Operations Management*, 6(2), pp 159–177.

Benjaafar, S. and Kim, J. S. (2004). When does higher demand variability lead to lower safety stocks. *Management Science*, In review.

Benjaafar, S., Kim, J. S. and Vishwanadham, N. (2004). On the effect of product variety in production-inventory systems. *Annals of Operations Research*, 126, pp 71–101.

Bobbio, A., Horváth, A., Scarpa, M. and Telek, M. (2003). Acyclic discrete phase type distributions: Properties and a parameter estimation algorithm. *Performance Evaluation*, 54(1), pp 1–32.

Bobbio, A., Horváth, A. and Telek, M. (2004a). 3-moments matching with minimal, positive acyclic phase-type distributions. *Research report.* Dept of Telecommunication, Technical University of Budapest.

Bobbio, A., Horváth, A. and Telek, M. (2004b). The scale factor: a new degree of freedom in phase type approximation. *Performance Evaluation*, 56(1-4), pp 121–144.

Chopra, S., Reinhardt, G. and Dada, M. (2004). The effect of lead time uncertainty on safety stocks. *Decision Sciences*, 35(1), pp 1–24.

Disney, S. M., Farasyn, I., Lambrecht, M. R., Towill, D. R. and Van de Velde, W. (2004). Dampening variability by using smoothing replenishment rules. *Working Paper.*

Graves, S. C. (1999). A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management*, 1(1), pp 50–61.

Hopp, W. J. and Spearman, M. L. (2001). *Factory Physics.* 2nd edn. Irwin, McGraw-Hill.

Horváth, A. and Telek, M. (2002). PhFit: A general phase-type fitting tool. *Proceedings of Performance TOOLS 2002.* London, UK.

Jemaï, Z. and Karaesmen, F. (2004). The influence of demand variability on the performance of a make-to-stock queue. *European Journal of Operational Research*, In Press.

Karaesmen, F., Buzacott, J. A. and Dallery, Y. (2002). Integrating advance order information in make-to-stock production systems. *IIE Transactions*, 34(8), pp 649–662.

Karaesmen, F., Liberopoulos, G. and Dallery, Y. (2004). The value of advance demand information in production-inventory systems. *Annals of Operations Research*, 126, pp 135–157.

Karmarkar, U. S. (1993). Manufacturing lead times, order release and capacity loading. *in* S. C. Graves, A. H. G. Rinnooy Kan and P. H. Zipkin (eds), *Logistics of Production and Inventory.* Vol. 4 of *Handbooks in Operations Research and Management Science.* Elsevier Science Publishers B.V. pp 287–329.

Kim, J. S. and Benjaafar, S. (2002). Extended abstract: On the benefits of inventory pooling in production-inventory systems. *Manufacturing and Service Operations Management Journal*, 4(1), pp 12–16.

Kleinrock, L. (1975). *Queueing Systems, Volume I: Theory.* Wiley. New York.

Latouche, G. and Ramaswami, V. (1999). *Introduction to Matrix Analytic Methods and stochastic modeling.* SIAM. Philadelphia.

Lu, Y., Song, J.-S. and Yao, D. D. (2003). Order fill rate, leadtime variability, and advance demand information in an assemble-to-order system. *Operations Research*, 51(2), pp 292–308.

Neuts, M. (1981). *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach.* John Hopkins University Press.

Neuts, M. (1989). *Structured Stochastic Matrices of M/G/1 type and their applications.* Marcel Dekker, Inc. New York and Basel.

Ramaswami, V. (1988). Nonlinear matrix equations in applied probability - solution techniques and open problems. *SIAM review*, 30(2), pp 256–263.

Ross, S. M. (1983). *Stochastic Processes.* John Wiley & Sons, New York.

Silver, E. A., Pyke, D. F. and Peterson, R. (1998). *Inventory Management and Production Planning and Scheduling.* 3rd edn. John Wiley & Sons, New York.

Song, J. S. (1994a). The effect of lead-time uncertainty in a simple stochastic inventory model. *Management Science*, 40(5), pp 603–613.

Song, J. S. (1994b). Understanding the lead-time effects in stochastic inventory systems with discounted costs. *Operations Research Letters*, 15, pp 85–93.

Song, J. S., Xu, S. H. and Liu, B. (1999). Order-fulfillment performance measures in an assemble-to-order system with stochastic lead times. *Operations Research*, 47(1), pp 131–149.

Song, J. S. and Yao, D. D. (2002). Performance analysis and optimization of assemble-to-order systems with random lead times. *Operations Research*, 50(5), pp 889–903.

Telek, M. (2000). Minimal coefficient of variation of discrete phase type distributions. *3rd International Conference on Matrix Analytic Methods in Stochastic Models.* Notable Publications Inc.. Leuven, Belgium. pp 391–400.

Telek, M. and Heindl, A. (2002). Matching moments for acyclic discrete and continuous phase-type distributions of second order. *International Journal of Simulation Systems, Science & Technology, Special issue on Analytical & Stochastic Modelling Techniques*, 3(3-4).

Thonemann, U. W. and Bradley, J. R. (2002). The effect of product variety on supply-chain performance. *European Journal of Operational Research*, 143(3), pp 548–569.

Vendemia, W. G., Patuwo, B. E. and Hung, M. S. (1995). Evaluation of lead time in production/inventory systems with non-stationary stochastic demand. *Journal of the operational research society*, 46, pp 221–233.

Zipkin, P. H. (2000). *Foundations of Inventory Management.* McGraw-Hill. New York.